Turning Binary Large-margin Bounds into Multi-class Bounds

Ürün Dogan DOGANUDB@MATH.UNI-POTSDAM.DE
Institut für Mathematik, Universität Potsdam, Germany

Tobias Glasmachers TOBIAS.GLASMACHERS@INI.RUHR-UNI-BOCHUM.DE
Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany

Christian Igel IGEL@DIKU.DK
Department of Computer Science, University of Copenhagen, Denmark

1. Introduction

The theory of generalization bounds for multi-class support vector machines (multi-class SVMs, Vapnik 1998; Weston & Watkins 1999; Bredensteiner & Bennett 1999; Crammer & Singer 2002; Lee et al. 2004) follows the route already paved by the analysis of binary large-margin classifiers (e.g. Guermeur, 2007; 2010). We link the analysis of binary and multi-class large margin classifiers explicitly by presenting a straightforward technique to generalize bounds for binary learning machines to the multi-class case.

2. Basic definitions

Given training data $S = \big((x_1, y_1), \ldots, (x_\ell, y_\ell)\big) \in \big(X \times Y\big)^\ell$ sampled i.i.d. from a fixed distribution P over an input space X and a finite label space Y, multi-class SVMs (including the one-versus-all method) learn a hypothesis $h : X \to Y$ of the form

$$x \mapsto \arg\max_{c \in Y} \big[\langle w_c, \phi(x)\rangle + b_c\big] \ , \qquad (1)$$

where $\phi : X \to \mathcal{H}$ is a feature map into an inner product space $\mathcal{H}$, $w_1, \ldots, w_d \in \mathcal{H}$ are class-wise weight vectors, and $b_1, \ldots, b_d \in \mathbb{R}$ are class-wise bias/offset values. We denote the risk of a hypothesis h w.r.t the 0-1 loss (i.e., the probability of misclassification) by $\mathcal{R}(h)$.

For a binary classifier based on thresholding a real-valued function $f^{\text{bin}} : X \to \mathbb{R}$ at zero, the empirical risk based on the hinge loss $L^{\text{hinge}}(f^{\text{bin}}(x), y) = \max\{0, 1 - y \cdot f^{\text{bin}}(x)\}$ is given by:

$$\mathcal{R}_S^{\text{hinge}}\left(f^{\text{bin}}\right) = \frac{1}{\ell} \sum_{i=1}^{\ell} L^{\text{hinge}}\left(f^{\text{bin}}(x_i), y_i\right)$$

One way of extending this definition to the multi-class case is by measuring the empirical risk with the sum loss

$$L^{\text{sum}}(f(x), y) = \sum_{c \in Y \setminus \{y\}} L^{\text{hinge}}\left(\frac{1}{2}\big(f_y(x) - f_c(x)\big), 2\delta_{y,c} - 1\right).$$

The corresponding empirical risk is denoted by $\mathcal{R}_S^{\text{sum}}$.

3. Main result

Our analysis relies on the basic insight that there are $d - 1$ distinct possible mistakes per example (x, y), namely preferring class $c \in Y \setminus \{y\}$ over the true class y. Each of these mistakes corresponds to one binary problem (having a decision function with weight vector $w_y - w_c$) indicating the specific mistake. One of these mistakes is sufficient for wrong classification, and no "binary" mistake at all implies correct classification. Then, a union bound over all mistakes gives the multi-class generalization result based on established bounds for binary classifiers.

Let us assume that we have a bound of the following generic form: With probability $1 - \delta$ over randomly drawn training sets S of size ℓ the risk $\mathcal{R}(f^{\text{bin}})$ of a binary classifier derived from a function $f^{\text{bin}} \in \mathbb{F}^{\text{bin}}$ is bounded by

$$\mathcal{R}(f^{\text{bin}}) \leq B^{\text{bin}}\left(\ell, \mathcal{R}_S^{\text{hinge}}(f^{\text{bin}}), \mathcal{C}(\mathbb{F}^{\text{bin}}), \delta\right) \ ,$$

where $\mathbb{F}^{\text{bin}}$ is a space of functions $X \to \mathbb{R}$. The function $\mathcal{C}$ measures the complexity of the function class $\mathbb{F}^{\text{bin}}$ in a possibly data-dependent manner (i.e., it may implicitly depend on properties of the training data, typically in terms of the kernel Gram matrix).

Then we have:

Theorem 1. *Given the aforementioned binary bound it holds: With probability $1-\delta$ over randomly drawn training sets $S \in \left(X \times \{1,\ldots,d\}\right)^{\ell}$, the risk $\mathcal{R}(f)$ of a multi-class classifier derived from the function $f = (f_1,\ldots,f_d) : X \to \mathbb{R}^d$, $f \in \mathbb{F}$, using the decision rule (1) is bounded by:*

$$\mathcal{R}(f) \leq \sum_{1\leq c<e\leq d} \left(\frac{\ell^{(c,e)}}{\ell} + \frac{1}{\sqrt{\ell}} \sqrt{\frac{\log(d(d-1)) - \log\delta}{2}} \right)$$
$$\cdot B^{bin}\left(\ell^{(c,e)}, \mathcal{R}^{hinge}_{S^{(c,e)}}\left(\frac{1}{2}(f_c - f_e)\right), \mathcal{C}(\mathbb{F}^{(c,e)}), \frac{\delta}{d(d-1)} \right)$$

Here $S^{(c,e)} = \left\{ (x,y) \in S \,\middle|\, y \in \{c,e\} \right\}$ is the training set restricted to examples of classes c and e, $\ell^{(c,e)} = \left|S^{(c,e)}\right|$ denotes its cardinality, and the pairwise binary function classes are defined as

$$\mathbb{F}^{(c,e)} = \left\{ \frac{1}{2}(f_c - f_e) \,\middle|\, f = (f_1,\ldots,f_d) \in \mathbb{F} \right\} .$$

4. Sample application

As an example, we apply our result to a simple textbook generalization bound for binary machines based on the Rademacher complexity. The underlying binary bound can be derived, for instance, following the proof of Theorem 4.17 by Shawe-Taylor & Cristianini (2004). We get:

Corollary 1. *Let $S \in \left(X \times \{1,\ldots,d\}\right)^{\ell}$ be a training set. Fix $\rho > 0$, and let $\mathbb{F}_\rho$ be the class of $\mathbb{R}^d$-valued functions in a kernel-defined feature space with semi-norm at most $1/\rho$ w.r.t. the semi-norm $\|f\| = \max\left\{ \frac{1}{2}\|f_c - f_e\| \,\middle|\, 1 \leq c < e \leq d \right\}$. With probability $1-\delta$ over randomly drawn training sets $S \in \left(X \times \{1,\ldots,d\}\right)^{\ell}$, the risk $\mathcal{R}(f)$ of a multi-class classifier using the decision rule (1) is bounded by*

$$\mathcal{R}(f) \leq \sum_{1\leq c<e\leq d} \left(\frac{\ell^{(c,e)}}{\ell} + \frac{1}{\sqrt{\ell}} \sqrt{\frac{\log(d(d-1)) - \log\delta}{2}} \right)$$
$$\cdot \left[\mathcal{R}^{hinge}_{S^{(c,e)}}\left(\frac{1}{2}(f_c - f_e)\right) + \frac{4}{\ell^{(c,e)}\rho}\sqrt{\mathrm{tr}(K^{(c,e)})} + 3\sqrt{\frac{\log(2d(d-1)/\delta)}{2\ell^{(c,e)}}} \right] ,$$

where $K^{(c,e)}$ denotes the $\ell^{(c,e)} \times \ell^{(c,e)}$ kernel matrix restricted to examples of classes c and e.

With $\mathrm{tr}(K^{(c,e)}) \leq \mathrm{tr}(K)$ this bound reads in soft-O notation (ignoring logarithmic terms)

$$\mathcal{R}(f) \in \tilde{O}\left(\frac{d(d-1)}{2} \left(\frac{4}{\rho \cdot \ell} \cdot \sqrt{\mathrm{tr}(K)} + \mathcal{R}^{\mathrm{sum}}_S(f) \right) \right) ,$$

with the same separation of complexity and empirical risk terms as in the binary bound.

5. Conclusion

The proposed way to extend generalization bounds for binary large-margin classifiers to large-margin multi-category classifiers is very simple, compared to taking all pairwise interactions between classes into account at once, and it has a number of advantageous properties. It is versatile and generic in the sense that it is applicable to basically every binary margin-based bound. Compared to the underlying bounds we pay the price of considering the worst case over $d(d-1)/2$ pairs of classes. However, also the state-of-the-art results obtained by Guermeur (2007) exhibit the same $\tilde{O}(d(d-1))$ scaling in the number of classes. In any case this term does not affect the asymptotic tightness of the bounds w.r.t. the number of samples. The same argument, put the other way round, implies that the asymptotic tightness of a bound for binary classification carries over one-to-one to the multi-class case. This implies that binary and multi-class learning have the same sample complexity.

References

Bredensteiner, E. J. and Bennett, K. P. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1):53–79, 1999.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2002.

Guermeur, Y. VC theory for large margin multi-category classifiers. *Journal of Machine Learning Research*, 8: 2551–2594, 2007.

Guermeur, Y. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, Application to multi-class SVMs. *Communications in Statistics: Theory and Methods*, 39(3):543–557, 2010.

Lee, Y., Lin, Y., and Wahba, G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465): 67–82, 2004.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Vapnik, V. *Statistical Learning Theory*. John Wiley and Sons, 1998.

Weston, J. and Watkins, C. Support vector machines for multi-class pattern recognition. In Verleysen, M. (ed.), *Proceedings of the Seventh European Symposium On Artificial Neural Networks (ESANN)*, pp. 219–224. Evere, Belgium: d-side publications, 1999.